CHAPTER 3: ITEM RATING WORKSHOPS

Two panels of educators were convened (one in northern California and one in southern California) to help evaluate items being developed for California's new High School Exit Exam (HSEE). Each panel participated in an item review workshop designed to provide two types of information: alignment of the field test items to California's Content Standards (California Department of Education, 1999d, 1999e) and alignment of the curriculum being taught in the classroom with each of the field test items.

Item Rating Procedures

Panel Members

The panel members were recruited from our longitudinal evaluation of 24 districts selected to be representative of the state sample as described in Chapter 1. We asked districts to identify individuals who were highly knowledgeable about the district's language arts or mathematics curriculum and instruction and were currently serving in either a teaching role or a district curriculum specialist role. The rating workshops were held on Saturdays and because of time and cost constraints, we were not able to include staff from all districts. We ended up with 44 panelists representing 13 of the 24 districts. Large, medium and small districts, as defined on the basis of the number of 10th grade students in our sampling procedures, were represented in the workshops as were a roughly equal number of low-ELL, high-ELL, low-mathematics, and high-mathematics districts. Large districts and districts with low 1999 STAR Math means were slightly underrepresented, but there were at least two districts from each category.

For the standards-alignment ratings, we expected general agreement across participants, regardless of the districts they came from. For the curriculum-alignment ratings, we did expect differences across districts and thus sought to be sure that different types of districts were included in our workshops. For both ratings, the primary focus was on differences among items. We were not yet trying to make inferences about differences among districts and so exact representation of all districts in the state was not a primary goal.

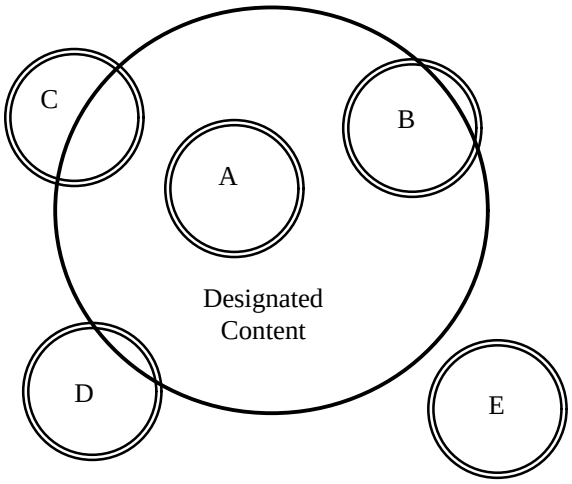
Standards Alignment Ratings

Mathematics or English Language Arts Content Standards[5] "Blueprints" for reading and mathematics have been recommended by the High School Exit Examination Standards Panel; they specify which particular components of each content strand should be included in the exam. Educators in the item review workshops judged the extent to which the items measure their intended content as designated by the item developers.

No single item can completely cover its targeted content. Using Venn diagrams, any given item can be represented as a small circle that is placed in some overlapping relationship to a larger circle representing the item's target content. Five placements are

[5] Reading and Mathematics are "domains," which are divided into "strands," which are further divided into "substrands," which are finally divided into "standards" (CDE, 1999). Items target "standards" in the blueprints.

depicted in Figure 3.1, representing five potential relationships between an item and its designated content.



Items A and B both show good content match. Item A is completely related to the content. For Item B, only a small portion of the item is outside the designated content area, perhaps indicating that it has some necessary language/communication components to convey the mathematical problem.

Items D and E both show clearly unacceptable content match. These items measure something other than designated content.

Figure 3.1 Five theoretical relationships between test items and their designated content.

Item C shows a partial, ambiguous match. This kind of item includes competencies or other factors (beyond necessary language/communication skills) in addition to the designated content. These items are "iffy" in the sense that psychometrically their overlap with the content could allow them to contribute valid test information, but their extraneous content could cause problems. These will also be "iffy" in the minds of the judges about whether or not they have a sufficient content match to be useful in the way intended by the test blueprint. For each field test item, panel members indicated whether each item was more like A, B, C, D, or E.

The items for Mathematics Reasoning (MR) represent a special case. Each of the MR items is obviously intended to assess mathematical reasoning and the reasoning problem for each item is intended to target one of the other mathematics strands. For example, an item may be intended to assess the MR strand "make and test conjectures by using both inductive and deductive reasoning" and attempt to do so by presenting a measurement and geometry problem. For such an item, raters made the additional judgment about whether or not the reasoning requirement involves the designated measurement and geometry standard.

Standards alignment was analyzed in terms of the proportion of items rated as A, B, C, D, and E (recoded to 5, 4, 3, 2, and 1 respectively) by content standard, by substrand, by strand, and by domain. A- and B-rated items clearly signal that, in the opinion of skilled educators, the item covers the intended content. To the extent that high proportions of these items are included in the field test, the potential for content validity of the assessment will be enhanced. C-rated items are more problematic, depending on the sources of extraneous content in the items. Some C-rated items could be used in an operational test if, as a set, they do not introduce systematic bias. High proportions of C-rated items signal the need to closely examine differential item functioning when field test results become available. High proportions of D- and E-rated items signal clear problems with test content or potential

balance in content, at least from the perspective of the panel members and their interpretation of the content frameworks.

Curriculum Alignment Ratings

Although panel members represented only a small proportion of the state, they were drawn from schools that were carefully selected to be representative of the state as a whole. Therefore, the panels provided an initial appraisal of the extent to which the content represented by the field test items is being taught to California students. The information is intended as a baseline against which we will assess changes in the alignment of curriculum toward the standards embodied in the HSEE as it is implemented over the next several years.

For each item, panel members were asked two questions:

1. What percent of your school's 10th grade students have been provided sufficient instruction to correctly answer this item?
2. What percent of your school's 12th grade students have been provided sufficient instruction to correctly answer this item?

The panelists provided answers on the following scale:

(4) More than 95%
(3) 75% to 94%
(2) 50% to 74%
(1) fewer than 50%

In analyzing curriculum alignment, we focused on the ratings for 10th grade students, since all students will be required to take the HSEE in either the 9th or 10th grade. We flagged items with an average rating below 2.5 on this 4-point scale as this implied that fewer than 75% of students (across all districts) had been provided sufficient instruction to correctly answer the item. Put another way, for items with these low ratings, more than 25% of all 10th grade students may not have had an opportunity to learn the material tested by the item.

Rating Booklets

For each subject, four field test forms with 99 to 102 items each were prepared. We divided each of these forms into two rating booklets of about 50 items each and asked panelists to rate as many of these booklets as possible. Booklets were assigned to panelists in a spiraled fashion so that we had approximately the same number of raters for each booklet and thus for each item.

Rater Training and Calibration

After raters were presented descriptions of their rating tasks and allowed to ask questions, all raters rated the same set of approximately five reading or math items. The group then reviewed the ratings and discussed rationales. This review process helped the raters better understand the rating task by clarifying differences, particularly between B and C rated items.

Group Discussion

In addition to the rating tasks, educators also engaged in a discussion period at the end of the day, after they had rated a sizable number of items. They were asked first to rate the proportion of their districts' students who they think would have been exposed to each content strand as characterized by the test items. A general discussion about their districts' plans or need for plans to facilitate students' preparation for the exit exam immediately followed.

Results

Standards Alignment

The standards alignment ratings used a 5-point scale, with items rated as 4 and 5 (denoted B and A in the discussion above) classified as having good alignment. The mean ratings for each items had an estimated reliability of .69 for each subject (.693 for ELA ratings and .687 for mathematics ratings). While not high enough to use as a basis for making important decisions about individual items, these values indicate a general agreement among the ratings as to differences in standards-alignment across items. The standard error of the item means was .29 for ELA and .38 for math. At this level of accuracy, differences of about three-quarters of a point were statistically significant. Our primary concern, however, was the average of the ratings for all items and for major groupings of items, and not with rating differences for specific items. The accuracy of these statistics was much greater, with standard errors less than .15 for the overall average.

Tables 3.1 and 3.2 show the results for the alignment of the items to their targeted standards. In general, the judges agreed closely with prior panels of experts who had screened these items. Nearly 80% of the ELA items and well over 90% of the mathematics items were judged to be "on target." Within ELA, the greatest concern was with the Writing Application items. These were all extended constructed response (essay) items, but scoring rubrics and anchor papers were not yet available for these items. These items will need to be re-examined when more information on scoring procedures becomes available. There were also concerns about a number of reading comprehension items, particularly items having to do with synthesis or comparison.

For mathematics, the mathematical reasoning category was the most problematic, followed by statistics, data analysis and probability at the Grade 7 level. We recommend more careful analysis of the mathematical reasoning standards and how they are assessed. The only standard with more than two low-rated items was "Make and test conjectures by using both inductive and deductive reasoning." As with reading, it appears that it is more difficult to write items for standards involving higher levels of synthesis and analysis. In all, however, only 35 of the 396 mathematics items were flagged for potential problems with alignment to the targeted standard.

Table 3.1 Total Items and Number With Low-Standards Alignment Ratings for Each Major Language Arts Content Category

SUBJECT/STRANDS	TARGET NO. ITEMS PER TEST FORM	NO. ITEMS IN THE FIELD TEST	NO. WITH LOW ALIGNMENT RATINGS*	PERCENT FLAGGED
Reading Vocabulary:				
Word Analysis, Fluency, and Vocabulary Development (RV)	10	42	3	7.1%
Reading Comprehension:				
Focus on Informational Materials (RI)	30	121	43	35.5%
Literary Response and Analysis (RL)	30	68	16	23.5%
Writing:				
Writing Strategies (WS)	12	58	7	12.1%
Written and Oral English Language Conventions (WC)	18	61	2	3.3%
Writing Applications (WA)	2	12	12	100.0%
TOTAL ALL ITEMS	**102**	**362**	**83**	**22.9%**

* Note: Items were flagged if the average standards-alignment rating was less than 3.5 indicating significant differences between item content and the content of the targeted standard.

Table 3.2 Total Items and Number With Low Standards Alignment Ratings for Each Major Mathematics Content Category

CONTENT AREA (STRAND)	TARGET NO. ITEMS PER TEST FORM	NO. ITEMS IN THE FIELD TEST	NO. WITH LOW ALIGNMENT RATINGS*	PERCENT FLAGGED
Statistics, Data Analysis, and Probability (Grade 6)	6	41	1	2.4%
Statistics, Data Analysis, and Probability (Grade 7)	8	23	4	17.4%
Number Sense	14	63	3	4.8%
Algebra and Functions	17	77	8	4.8%
Measurement and Geometry	20	80	5	6.3%
Mathematical Reasoning	8	43	9	20.9%
Algebra 1	26	69	5	7.2%
TOTAL ALL ITEMS	**99**	**396**	**35**	**8.8%**

* Note: Items were flagged if the average standards-alignment rating was less than 3.5 indicating significant differences between item content and the content of the targeted standard.

Curriculum Alignment

Tables 3.3 and 3.5 show the numbers of items with relatively low curriculum alignment ratings for ELA and math respectively. We relied on the 10th grade ratings, since all students will be required to take the HSEE in either 9th or 10th grade. We used a cutoff of 2.5 to separate items where more than 25% of the students were judged not to have been prepared to answer the item. In analyzing the curriculum-alignment ratings, we excluded items that had been flagged for low standards-alignment. The excluded items were judged to be poor measures of the target standards and so the alignment of the item to the district's curriculum was not relevant.

For both ELA and math, our panelists believed that, for a majority of the items, there are a significant number of students who have not had the opportunity to learn the skills tapped by these items. The ELA panelists, in particular, judged that over 90% of the standards-aligned items could cause problems for their 10th grade students. Writing Conventions was the only category where fewer than 90% of the items were flagged.

For Mathematics, the panelists rated just over half of the items as having potential curriculum-alignment problems. There was considerable variation across the different math content categories in these ratings. Nearly 80% of the Algebra 1 items and 64% of the Algebra and Functions items were flagged. The percentage of items flagged in the other content categories was significantly less.

Since each panelist rated alignment to the curriculum of a different district, we could not use inter-rater agreement as a measure of the accuracy of the curriculum-alignment ratings. Instead, we used information on the difficulty of the test items in the field test (see Chapter 4) as an indicator of the validity of the ratings. Test questions may be difficult for several reasons, one of which is that a significant number of students have not been taught the content measured by the item. We would expect that, on average, items that were not aligned to the curriculum of a significant number of students would be more difficult than items that were aligned to the curriculum of most students. Table 3.5 shows that this was, in fact, the case for both the ELA and the math item reviews. The difference in passing rates between high and low alignment items was nearly 30 percentage points for both subjects.

Table 3.3 Total Items and Number With Low Curriculum Alignment Ratings for each Major Language Arts Content Category

SUBJECT/STRANDS	NO. ITEMS IN THE FIELD TEST	NO. ALIGNED TO TARGET STANDARD	NO. WITH LOW CURRICULUM ALIGNMENT*	PERCENT FLAGGED
Reading Vocabulary:				
Word Analysis, Fluency, and Systematic Vocabulary Development	42	39	36	92.3%
Reading Comprehension:				
Focus on Informational Materials	121	78	74	94.9%
Literary Response and Analysis	68	52	48	92.3%
Writing:				
Writing Strategies	58	51	47	92.2%
Written and Oral English Language Conventions	61	59	48	81.4%
Writing Applications	12	0	0	---
TOTAL ITEMS	**362**	**279**	**253**	**90.7%**

* Note: Items were flagged if the average for the 10th grade student ratings was less than 2.5 indicating that more than 25% of the 10th graders had not received sufficient instruction to prepare them to answer the item correctly.

Table 3.4 Total Items and Number With Low Curriculum Alignment Ratings for Each Major Math Content Category

CALIFORNIA CONTENT STANDARD	NO. ITEMS IN THE FIELD TEST	NO. ALIGNED TO TARGET STANDARD	NO. WITH LOW CURRICULUM ALIGNMENT*	PERCENT FLAGGED
Statistics, Data Analysis, and Probability (Grade 6)	41	40	15	37.5%
Statistics, Data Analysis, and Probability (Grade 7)	23	19	7	36.8%
Number Sense	63	60	18	30.0%
Algebra and Functions	77	69	44	63.8%
Measurement and Geometry	80	75	36	48.0%
Mathematical Reasoning	43	34	14	41.2%
Algebra 1	69	64	51	79.7%
TOTAL ITEMS	**396**	**361**	**185**	**51.2%**

* Note: Items were flagged if the average for the 10th grade student ratings was less than 2.5 indicating that more than 25% of the 10th graders had not received sufficient instruction to prepare them to answer the item correctly.

Table 3.5 Relationship of Item Difficulty to Curriculum Alignment Ratings

SUBJECT	CURRICULUM ALIGNMENT	PERCENT ITEMS WITH MORE THAN 50% PASSING IN THE FIELD TEST
Language Arts	High	93.9%
	Low	67.7%
Mathematics	High	59.4%
	Low	30.9%

Conclusions

The results of our analyses confirmed that most of the items developed for the field test are reasonably aligned to the state standards. Even though our panelists were not necessarily experts in these standards, they reached reasonable agreement among themselves on differences in alignment across the pool of field test items and they agreed as well with the HSEE Panel who had examined the alignment of these items. A modest number of items were flagged for further review. Nearly all of these items had standards-alignment ratings near the margin of acceptability.

With respect to the alignment of the test items to the current curriculum in different districts, our results suggest that curriculum specialists need to bring the curriculum into alignment with targeted standards and items tested on the HSEE. Current results are only baseline information. The State Board has not yet adopted these standards and the students who will have to meet the standards have not yet reached high school (they have now completed the 8th grade). It will be important to repeat this exercise next spring to see whether the alignment of the curriculum to the items and hence to the standards is increasing.